Statistical Mechanics of Dictionary Learning

Ayaka Sakata[1] and Yoshiyuki Kabashima[1]

[1] *Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan.*

PACS 89.20.Ff – Computer science and technology
PACS 75.10.Nr – Spin-glass and other random models

Abstract – Finding a basis matrix (dictionary) by which objective signals are represented sparsely is of major relevance in various scientific and technological fields. We consider a problem to learn a dictionary from a set of training signals. We employ techniques of statistical mechanics of disordered systems to evaluate the size of the training set necessary to typically succeed in the dictionary learning. The results indicate that the necessary size is much smaller than previously estimated, which theoretically supports and/or encourages the use of dictionary learning in practical situations.

Introduction. – In various fields of science and technology, such as earth observation, astronomy, medicine, civil engineering, materials science, and in compiling image databases [1], it has a major relevance to recover original signals from deficient signals obtained by limited number of measurements. The Nyquist-Shannon sampling theorem [2] provides the necessary and sufficient number of measurements for recovering arbitrary band-limited signals. However, techniques based on this theorem sometimes do not match restrictions and/or demands of today's front-line applications [3, 4], and much effort is still being made to find more efficient methodologies.

The concept of sparse representations has recently drawn great attention in such research. Many real world signals such as natural images are represented *sparsely* in Fourier/wavelet domains; namely, many components vanish or are negligibly small in amplitude when the signals are represented by Fourier/wavelet bases. This empirical property is exploited in the signal recovery paradigm of compressed sensing (CS) enabling the recovery of sparse signals from much fewer measurements than those the sampling theorem estimates [5–10].

However, the effectiveness of CS relies considerably on the assumption that a basis by which the objective signals look sparse is known in advance. Therefore, in applying CS to general signals of interest, whose bases for sparse representation are unknown, the primary task to accomplish is to identify an appropriate basis (dictionary) for the sparse representation from an available set of training signals. This is often termed *dictionary learning (DL)* [11–13].

Let us denote the training set of M-dimensional signals as an $M \times P$ matrix $\boldsymbol{Y} = \{Y_{\mu l}\}$, where each column vector $\boldsymbol{Y}_l$ represents a sample signal and P is the number of the samples. In a simple scenario, DL is formulated as a problem to find a pair of an $M \times N$ matrix (dictionary) $\boldsymbol{D} = \{D_{\mu i}\}$ and an $N \times P$ sparse matrix $\boldsymbol{X} = \{X_{il}\}$ such that $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$ holds. By DL, the characteristics/trends underlying $\{\boldsymbol{Y}_l\}$ are extracted into $\boldsymbol{D}$, and $\boldsymbol{Y}_l$ can be compactly represented as a superposition of a few dictionary columns, whose combination and strength are specified by the sparse matrix $\boldsymbol{X}$. DL suits not only efficient signal processing such as CS, but also extraction of non-trivial regularities from high-dimensional data. For instance, DL has been successfully applied to the facial image processing for the efficient storage of large databases, where standard algorithms fail [12, 14]. In this case, $\boldsymbol{Y}_l$ and $\boldsymbol{D}$ correspond to a facial image and a collection of patches of facial patterns learned by the P samples, respectively. A variant of DL has also been employed in gene expression analysis to estimate transcription factor activity $\boldsymbol{D}$ from gene expression data of a small size $\{\boldsymbol{Y}_l\}$ [15].

An important question of DL is how large a sample size P is necessary to uniquely identify an appropriate dictionary $\boldsymbol{D}$, because the ambiguity of the dictionary is fatal in use for signal/data analysis after learning. As the first answer to this question, an earlier study based on linear algebra showed that when the training set $\boldsymbol{Y}$ is generated by a pair of matrices $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$ (planted solution) as $\boldsymbol{Y} = \boldsymbol{D}^0\boldsymbol{X}^0$, one can perfectly learn these as a unique solution except for the ambiguities of signs and permutations of matrix elements if $P > P_{\rm c} = (k+1)_N C_k$ and k is sufficiently small, where k is the number of non-zero elements in each column of $\boldsymbol{X}^0$ [16]. This result is significant as it is the first proof that guarantees the learnability with a finite size sample set for DL. However, the estimate of $P_{\rm c}$ is supposed to enable a considerable improvement; the authors of [16] speculated that $P_{\rm c}$ could be reduced substantially to $O(N^2)$ or even smaller, although providing a mathematical proof was technically difficult. The improvement of the estimation $P_{\rm c}$ is practically significant because it will lead to considerable reduction of necessary cost for DL in terms of both sample and computational complexities.

In this Letter, we take an alternative approach to estimating $P_{\rm c}$. Specifically, we examine the typical behavior of DL using the replica method in the limit of $N, M, P \to \infty$. The main result of our analysis is that the planted solution is typically learnable by $O(N)$ training samples if negligible mean square errors per element are allowed and M/N is sufficiently large. This theoretically supports and/or encourages the employment of DL in practical applications.

Problem setting. – We focus on the learning strategy

$$\min_{\boldsymbol{D},\boldsymbol{X}} ||\boldsymbol{Y}(=N^{-1/2}\boldsymbol{D}^0\boldsymbol{X}^0) - N^{-1/2}\boldsymbol{D}\boldsymbol{X}||^2$$
$$\text{subj. to } ||\boldsymbol{D}||^2 = MN,\ ||\boldsymbol{X}||_0 = NP\theta \qquad (1)$$

[11, 12, 16–19], where $||\boldsymbol{A}||^2 = \sum_{\mu l} A_{\mu l}^2$ for a matrix $\boldsymbol{A} = \{A_{\mu l}\}$, and $||\boldsymbol{A}||_0$ represents the number of non-zero elements of $\boldsymbol{A}$. The parameter $\theta \in [0,1]$ denotes the rate of non-zero elements assumed by the learner, and $N^{-1/2}$ is introduced for convenience in taking the large system limit.

For simplicity, we assume that $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$ of the planted solution are uniformly generated under the constraints of $||\boldsymbol{D}^0||^2 = MN$, $||\boldsymbol{X}^0||_0 = NP\rho$ and $||\boldsymbol{X}^0||^2 = NP\rho$. We consider that the correct non-zero density ρ can differ from θ for generality, but we assume $\rho \le \theta$; otherwise, the correct identification of $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$ is trivially impossible. The main goal of our study is to evaluate the critical sample ratio $\gamma_{\rm c} = P_{\rm c}/N$ above which the planted solution can be learned typically.

Statistical mechanics approach. – Partition function

$$Z_\beta(\boldsymbol{D}^0,\boldsymbol{X}^0) = \int d\boldsymbol{D}d\boldsymbol{X} \exp\left(-\frac{\beta}{2N}||\boldsymbol{D}\boldsymbol{X} - \boldsymbol{D}^0\boldsymbol{X}^0||^2\right)$$
$$\times\, \delta(||\boldsymbol{D}||^2 - NM)\delta(||\boldsymbol{X}||_0 - NP\theta) \qquad (2)$$

constitutes the basis of our approach since the minimized cost of eq. (1) can be identified with the zero temperature free energy $F = -\lim_{\beta\to\infty}\beta^{-1}\ln Z(\boldsymbol{D}^0,\boldsymbol{X}^0;\beta)$. This statistically fluctuates depending on $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$. However, as $N, M, P \to \infty$, one can expect that the *self-averaging* property is realized; i.e., the free energy density $N^{-2}F$ converges to the typical value $f \equiv N^{-2}[F]_0$ with probability unity, where $[\cdots]_0$ stands for the average with respect to $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$. Consequently, this property is also expected to hold for other relevant macroscopic variables of the solution of eq. (1), $\boldsymbol{D}^*$ and $\boldsymbol{X}^*$. Therefore, assessing f is the central issue in our analysis.

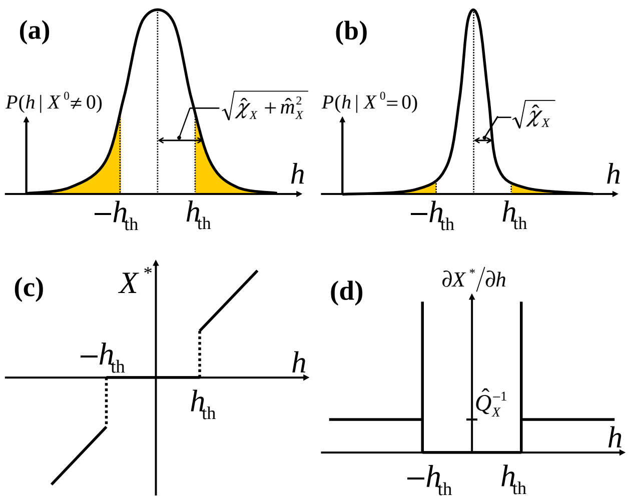

Fig. 1: (color online) (a) and (b) show distributions of local field h (a) $P(h|X^0 \neq 0)$ for $X^0 \neq 0$ and (b) $P(h|X^0 = 0)$ for $X^0 = 0$. (c) and (d) show X^* and $\partial X^*/\partial h$ as functions of h, respectively.

This assessment can be carried out systematically using the replica method [20] in the limit of $N \to \infty$ while keeping $\alpha = M/N \sim O(1)$ and $\gamma = P/N \sim O(1)$. Under the replica symmetric (RS) ansatz, where the solution space of eq. (1) is assumed to be composed of at most a few pure states [21], the free energy density is given as

$$f = \underset{\Omega,\hat{\Omega}}{\rm extr}\Big\{-\alpha\Big(\frac{\hat{Q}_D - \hat{\chi}_D\chi_D}{2} - \hat{m}_D m_D + \frac{\hat{\chi}_D + \hat{m}_D^2}{2\hat{Q}_D}\Big)$$
$$-\gamma\Big(\frac{\hat{Q}_X Q_X - \hat{\chi}_X\chi_X}{2} - \hat{m}_X m_X + \lambda\theta - \langle\langle\phi(h;\hat{Q}_X,\lambda)\rangle\rangle_h\Big)$$
$$+\frac{\alpha\gamma(Q_X - 2m_D m_X + \rho)}{2(1 + Q_X\chi_D + \chi_X)}\Big\}, \qquad (3)$$

where ${\rm extr}_{\Omega,\hat{\Omega}}\{\mathcal{G}(\Omega,\hat{\Omega})\}$ stands for the extremization of a function $\mathcal{G}(\Omega,\hat{\Omega})$ with respect to a set of macroscopic variables $\Omega \equiv \{\chi_D, m_D, Q_X, \chi_X, m_X\}$ and that of their conjugates $\hat{\Omega} \equiv \{\hat{Q}_D, \hat{\chi}_D, \hat{m}_D, \hat{Q}_X, \hat{\chi}_X, \hat{m}_X, \lambda\}$, and

$$\phi(h;\hat{Q}_X,\lambda) = \min_X \lim_{\epsilon\to+0}\Big\{\frac{\hat{Q}_X X^2}{2} - hX + \lambda|X|^\epsilon\Big\}. \qquad (4)$$

Notation $\langle\langle\cdots\rangle\rangle_h$ represents the average with respect to h according to the distribution $P(h) = \rho P(h|X^0 \neq 0) + (1-\rho)P(h|X^0 = 0)$, where $P(h|X^0 \neq 0)$ and $P(h|X^0 = 0)$ are given by zero-mean Gaussian distributions with variances $\hat{\chi}_X + \hat{m}_X^2$ and $\hat{\chi}_X$, respectively (Fig. 1(a),(b)). The details of the derivation of the free energy density are shown in Appendix.

Physical implications. – At the extremum of eq. (3), the relationships

$$m_D = \frac{1}{MN}[\mathrm{Tr}(\boldsymbol{D}^0)^{\mathrm{T}}\boldsymbol{D}^*]_0, \tag{5}$$

$$m_X = \frac{1}{NP}[\mathrm{Tr}(\boldsymbol{X}^0)^{\mathrm{T}}\boldsymbol{X}^*]_0, \tag{6}$$

$$Q_X = \frac{1}{NP}[\mathrm{Tr}(\boldsymbol{X}^*)^{\mathrm{T}}\boldsymbol{X}^*]_0 = \frac{1}{NP}[||\boldsymbol{X}^*||^2]_0 \tag{7}$$

hold, where T denotes the matrix transpose. These provide the mean square errors (per element), which measure the performance of DL, as

$$\epsilon_D \equiv \frac{1}{MN}[||\boldsymbol{D}^* - \boldsymbol{D}^0||^2]_0 = 2(1-m_D) \tag{8}$$

$$\epsilon_X \equiv (NP)^{-1}[||\boldsymbol{X}^* - \boldsymbol{X}^0||^2]_0 = \rho - 2m_X + Q_X. \tag{9}$$

The variables χ_D and χ_X physically mean the sensitivity of the estimates $\boldsymbol{D}^*$ and $\boldsymbol{X}^*$ when the cost of eq. (1) is linearly perturbed.

Eq.(4) represents the effective single-body minimization problem concerning an element of $\boldsymbol{X}$ that is statistically equivalent to eq. (1) [22]. Here, the randomness of $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$ is effectively replaced by the random local field h. The first and second terms of $P(h)$ correspond to the cases where an element of $\boldsymbol{X}^0$ is given as $X^0 \neq 0$ and $X^0 = 0$, respectively. Under a given h, the solution X^* that minimizes the cost of eq. (4) is offered as $X^* = h/\hat{Q}_X$ for $|h| > h_{\mathrm{th}} \equiv (2\hat{Q}_X\lambda)^{1/2}$ and 0 otherwise (Fig. 1 (c)). We refer to the cases of $|h| > h_{\mathrm{th}}$ and $|h| < h_{\mathrm{th}}$ as "active" and "inactive," respectively. When $X^0 \neq 0$, h is generated from a Gaussian distribution ($P(h|X^0 \neq 0)$) of zero-mean and variance $\hat{\chi}_X + \hat{m}_X^2$, and X^* is more likely to be active than when $X^0 = 0$, for which h is characterized by another zero-mean Gaussian ($P(h|X^0 = 0)$) of a smaller variance $\hat{\chi}_X$ (Fig. 1 (a),(b)). Therefore, one can expect that the hard-thresholding scheme based on h_{th} represents proper assignment of zero/non-zero elements in $\boldsymbol{X}^*$ so as to accurately estimate $\boldsymbol{X}^0$ and $\boldsymbol{D}^0$ if $\hat{m}_X$ is sufficiently large.

A distinctive feature of X^* is the divergence of the local susceptibility $\partial X^*/\partial h$ at "border" cases of $h = \pm h_{\mathrm{th}}$ (Fig. 1 (d)). This affects the increase in the effective degree of freedom (ratio) as follows: $\theta_{\mathrm{eff}} = \theta + \langle\langle h_{\mathrm{th}}\delta(|h| - h_{\mathrm{th}})\rangle\rangle_h$, whereas h_{th} is determined so as to satisfy $\theta = \int_{|h|>h_{\mathrm{th}}} dhP(h)$ indicating the sparsity condition $||\boldsymbol{X}||_0 = NP\theta$. The excess $\langle\langle h_{\mathrm{th}}\delta(|h| - h_{\mathrm{th}})\rangle\rangle_h$ is supposed to represent a combinatorial complexity for classifying each element of $\boldsymbol{X}^*$ that corresponds to the border case $|h| = h_{\mathrm{th}}$ into the active case, $|h| > h_{\mathrm{th}}$ and $\boldsymbol{X}^* \neq 0$, or the inactive case, $|h| < h_{\mathrm{th}}$ and $\boldsymbol{X} = 0$. The divergence of $\partial X^*/\partial h|_{h=\pm h_{\mathrm{th}}}$ is also accompanied by the instability of the RS solution against perturbations that break the replica symmetry [23]. The influence of this instability is discussed later.

Actual solutions. – We found two types of solutions; the first one is characterized by $m_D = 1$ and $Q_X = m_X = \rho$, while the second is characterized by $m_X = 0$ and $m_D = 0$. The former case provides $\epsilon_D = \epsilon_X = 0$ indicating the correct identification of $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$, and hence we call it the *success* solution. The latter is referred to as the *failure* solution since $m_D = 0$ and $m_X = 0$ indicate the complete failure of information extraction of $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$.

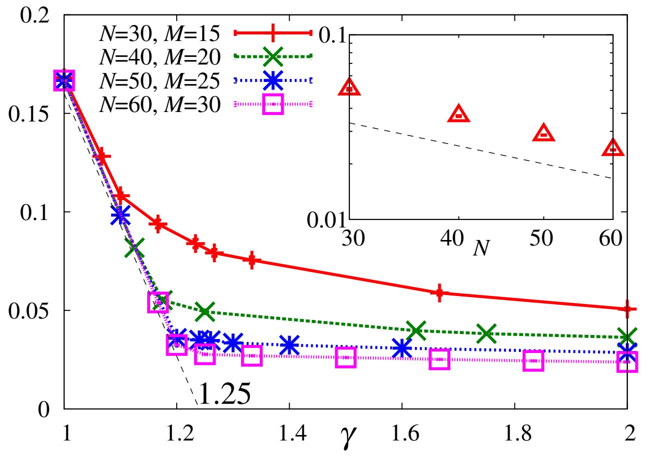

Fig. 2: (color online) γ-dependence of the ratio of marginal modes relative to N^2 at $\alpha = 0.5$ and $\rho = \theta = 0.1$. The behavior at $N \to \infty$ extrapolating from the results of finite N is denoted by the dashed line. Inset: N-dependence of the ratio of marginal modes for $\gamma = 2$. The dashed line stands for N^{-1} as a guide. Each marker represents the average of 100 experiments.

Success solution (**S**) exists when $\gamma > 1$ and

$$\alpha > \theta_{\mathrm{eff}}^{\mathrm{S}}(\theta, \rho) = \theta + (1-\rho)\sqrt{\frac{2}{\pi}}u\exp\left(-\frac{u^2}{2}\right) \tag{10}$$

hold, where $u = H^{-1}((\theta - \rho)/(2(1-\rho)))$ and $H^{-1}(x)$ is the inverse function of $H(x) = (2\pi)^{-1/2}\int_x^\infty dte^{-t^2/2}$. **S** is further classified into two cases depending on γ. For $\gamma > \gamma_{\mathrm{S}}$, where

$$\gamma_{\mathrm{S}}(\alpha, \theta, \rho) = \frac{\alpha}{\alpha - \theta_{\mathrm{eff}}^{\mathrm{S}}}, \tag{11}$$

χ_D and χ_X are finite. On the other hand, for $1 < \gamma < \gamma_{\mathrm{S}}$, χ_D and χ_X tend to infinity, keeping $R = \chi_D/\chi_X$ finite.

To physically interpret this classification, let us take a variation around $\boldsymbol{Y} = N^{-1/2}\boldsymbol{D}^0\boldsymbol{X}^0$, which yields

$$0 = \delta(\boldsymbol{D}\boldsymbol{X})|_{\boldsymbol{D}^0,\boldsymbol{X}^0} = \boldsymbol{D}^0\delta\boldsymbol{X} + \delta\boldsymbol{D}\boldsymbol{X}^0. \tag{12}$$

If $\delta\boldsymbol{D} = 0$ and $\delta\boldsymbol{X} = 0$ are the unique solution of eq. (12), the planted solution is locally stable. Otherwise, there are "marginal" modes along which the cost of eq. (1) does not increase locally, and the solution set forms a manifold. The number of constraints of eq. (12), MP, coincides with that of the degree of freedom of $\delta\boldsymbol{D}$ and $\delta\boldsymbol{X}$, $MN + NP\theta_{\mathrm{eff}}^{\mathrm{s}}$, at $P = \gamma_{\mathrm{S}}N$. Thus, the classification below/above γ_{S} corresponds to the change in the number of marginal modes around the planted solution.

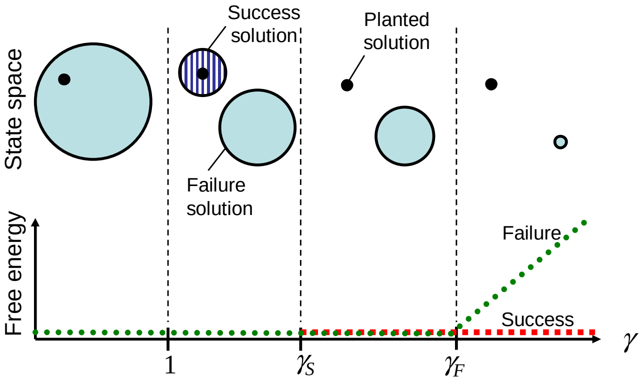

Fig. 3: (color online) Schematic pictures of γ-dependence of the phase space and free energy under RS assumption.

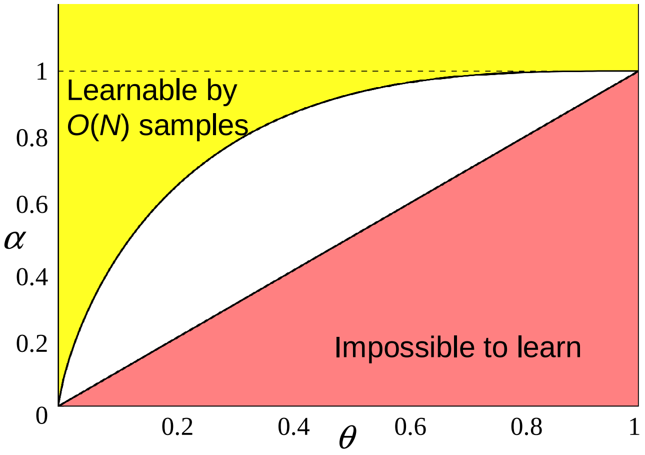

Fig. 4: (color online) Phase diagram on $\alpha-\theta$ plane.

To confirm the validity of this interpretation, we numerically evaluated the number of marginal modes of eq. (12) in the case of $\alpha=1/2$ and $\theta=\rho=0.1$, which is shown in Fig. 2. The assessment of $\gamma_{\rm s}$ when $\theta=\rho$ is conjectured to be exact since the effect of the border elements is negligible under this condition. Fig. 2 indicates that the number of marginal modes scales as $O(N^2)$ for $\gamma<\gamma_{\rm S}=1.25$, while it scales as $O(N)$, and the contribution of the marginal modes approaches zero, for $\gamma>\gamma_{\rm S}$ (inset). This result coincides with our theoretical assessment. At the same time, this implies that identifying the planted solution without any errors by eq. (1) is difficult as long as $\gamma\sim O(1)$, but the discrepancies per element caused by the marginal modes are negligibly small and could be allowed in many practical situations.

In the case of $\gamma<1$, for any $N\times P$ matrix $\boldsymbol{Z}$ of $||\boldsymbol{Z}||_0=NP\theta$, $\boldsymbol{X}^*=a\boldsymbol{Z}$ and $\boldsymbol{D}^*=a^{-1}\boldsymbol{Y}(\boldsymbol{Z}\boldsymbol{Z}^{\rm T})^{-1}\boldsymbol{Z}^{\rm T}$ minimize the cost of (1) to zero, where a is determined such that $||\boldsymbol{D}^*||^2=MN$. This implies that the set of solutions of eq. (1) spreads widely, and the weight of the planted solution is negligibly small in the state space. This may be why **S** disappears for $\gamma<1$.

Failure solution (**F**) exists for $\forall\gamma\geq 0$. If

$$\alpha<\theta_{\rm eff}^{\rm F}(\theta)=\theta+\sqrt{\frac{2}{\pi}}v\exp\left(-\frac{v^2}{2}\right) \tag{13}$$

where $v=H^{-1}(\theta/2)$ holds, **F** always offers $\chi_D,\chi_X\to\infty$ making the free energy f vanish. For $\alpha>\theta_{\rm eff}^{\rm F}$, on the other hand, χ_D and χ_X become finite implying that a single solution of eq. (1) is locally stable for most directions and offers $f>0$, if γ is greater than

$$\gamma_{\rm F}(\alpha,\theta)=\frac{\alpha}{(\alpha^{1/2}-(\theta_{\rm eff}^{\rm F})^{1/2})^2}. \tag{14}$$

The inequality $\theta_{\rm eff}^{\rm F}\geq\theta_{\rm eff}^{\rm S}$ always holds because the influence of the border elements for **F** is stronger than that for **S**, which leads to $\gamma_{\rm S}\leq\gamma_{\rm F}$.

Fig. 3 illustrates changes in state space that occur for sufficiently large α under the RS assumption. For $\gamma<1$, **F** is a unique solution. As γ increases, **S** appears at $\gamma=1$, and the number of marginal modes changes from $O(N^2)$ to $O(N)$ at $\gamma=\gamma_{\rm S}$. This implies that when negligibly small linear perturbations are added to the cost of eq. (1), the limits $\lim_{N\to\infty}\epsilon_D\sim 0$ and $\lim_{N\to\infty}\epsilon_X\sim 0$ still hold for **S** of $\gamma>\gamma_{\rm S}$ while they can be boosted to $O(1)$ for **S** of $\gamma<\gamma_{\rm S}$. For $\gamma<(\gamma_{\rm S}\leq)\gamma_{\rm F}$, **S** and **F** are degenerated providing $f=0$. However, at $\gamma=\gamma_{\rm F}$, **S** becomes thermodynamically dominant by keeping $f=0$, while **F** begins to have positive f. This means that the planted solution is typically learnable by $P>P_{\rm c}=N\gamma_{\rm F}\sim O(N)$ training samples if negligible mean square errors per element are allowed.

Fig. 4 plots the phase diagram on an $\alpha-\theta$ plane. The region above $\alpha=\theta_{\rm eff}^{\rm F}(\theta)$ (curve) represents the condition under which the planted solution is typically learnable by $O(N)$ training samples. DL is impossible in the region below $\alpha=\theta$ (straight line) because $\boldsymbol{X}^0$ cannot be correctly recovered even if $\boldsymbol{D}^0$ is known [7]. How the sample complexity scales with respect to N in the region $\theta<\alpha<\theta_{\rm eff}^{\rm F}(\theta)$ is beyond the scope of this Letter, but an interesting question nonetheless.

Summary and discussion. – In summary, we have assessed the size of training samples required for correctly learning a planted solution in DL using the replica method. Our analysis indicated that $O(N)$ samples, which are much fewer than estimated in an earlier study [16], are sufficient for learning a planted dictionary with allowance for negligible square discrepancies per element when the number of non-zero signals is sufficiently small compared to that of measurements.

It was shown that the identification of dictionary can be characterized as a phase transition with respect to the number of training samples. Our RS analysis probably does not describe the exact behavior of DL since the RS solutions are unstable against the replica symmetry breaking (RSB) disturbances. However, we still speculate that the RS estimate of γ_{F} serves as an upper bound of the correct critical ratio γ_{c}. This is because the free energy value of **F** assessed under the RSB ansatz should be greater than or equal to that of the RS solution due to the positivity constraint of the entropy of pure states (complexity) [24], whereas that of **S** is kept to vanish, which always yields a smaller estimate of γ_{F}.

Promising future research includes an extension of the current framework to more general situations such as noisy cases as well as refinement of the estimates of the critical ratios γ_{S} and γ_{F} taking RSB into account [25].

* * *

This work was partially supported by a Grant-in-Aid for JSPS Fellow (No. 23–4665) (AS) and KAKENHI No. 22300003 (YK).

Appendix: Derivation of eq. (3). – In general, the configurational average of the free energy density could be evaluated on the basis of the following formula:

$$f_\beta = -\frac{1}{\beta} \lim_{N\to\infty} \frac{1}{N^2} \lim_{n\to 0} \frac{\partial}{\partial n} \ln [Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0. \qquad (15)$$

Unfortunately, assessing $[Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0$ for $n \in \mathbb{R}$ in the mathematically rigorous manner is technically difficult, and this fact prohibits us from utilizing eq. (15) in practice. In the replica method, this difficulty is resolved by evaluating $N^{-2} \ln[Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0$ for $n \in \mathbb{N}$ as an analytic function of n first in the limit of $N \to \infty$, and taking the $n \to 0$ limit afterward with use of the obtained analytic function for $n \in \mathbb{R}$ as well.

More precisely, we evaluate $[Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0$ by averaging the right hand side of an identity

$$\begin{aligned} &Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0) = \\ &\int \prod_{a=1}^{n} \left\{ d\boldsymbol{D}^a \boldsymbol{X}^a \delta(||\boldsymbol{D}^a||^2 - NM) \delta(||\boldsymbol{X}^a||_0 - NP\theta) \right\} \\ &\qquad \times \exp\left(-\frac{\beta}{2N} \sum_{a=1}^{n} ||\boldsymbol{D}^a \boldsymbol{X}^a - \boldsymbol{D}^0 \boldsymbol{X}^0||^2 \right), \qquad (16) \end{aligned}$$

which is valid for only $n \in \mathbb{N}$, over the distributions of the planted solutions $\boldsymbol{D}^0$ and $\boldsymbol{X}^0$ that are given by

$$P_{D^0}(\boldsymbol{D}^0) = \frac{1}{\mathcal{N}_D} \delta(||\boldsymbol{D}^0||^2 - NM) \qquad (17)$$

and

$$P_{X^0}(\boldsymbol{X}^0) = \prod_{i,l} \left\{ (1-\rho)\delta(X_{i,l}) + \frac{\rho}{\sqrt{2\pi}} \exp\left(-\frac{X_{il}^2}{2} \right) \right\}, (18)$$

respectively, where $\mathcal{N}_D$ is the normalization constant. In performing the integrals of $2(n+1)$ variables $(\boldsymbol{D}^0, \{\boldsymbol{D}^a\})$ and $(\boldsymbol{X}^0, \{\boldsymbol{X}^a\})$ that come out in this evaluation, we insert trivial identities with respect to all combinations of replicas $(a, b = 0, 1, 2, \ldots, n)$,

$$1 = NM \int dq_D^{ab} \delta(\mathrm{Tr}(\boldsymbol{D}^a)^{\mathrm{T}} \boldsymbol{D}^b - NMq_D^{ab}), \qquad (19)$$

and

$$1 = NP \int dq_X^{ab} \delta(\mathrm{Tr}(\boldsymbol{X}^a)^{\mathrm{T}} \boldsymbol{X}^b - NPq_X^{ab}) \qquad (20)$$

to the integrand. Let us denote $\mathcal{Q}_D \equiv (q_D^{ab})$ and $\mathcal{Q}_X \equiv (q_X^{ab})$, and introduce two joint distributions

$$\begin{aligned} P_D(\{\boldsymbol{D}^a\}; \mathcal{Q}_D) &= \frac{P_{D^0}(\boldsymbol{D}^0)}{V_D(\mathcal{Q}_D)} \\ &\times \prod_{a=1}^{n} \delta(||\boldsymbol{D}^a||^2 - NM) \prod_{a<b} \delta(\mathrm{Tr}(\boldsymbol{D}^a)^{\mathrm{T}} \boldsymbol{D}^b - NMq_D^{ab}), \end{aligned} \qquad (21)$$

$$\begin{aligned} P_X(\{\boldsymbol{X}^a\}; \mathcal{Q}_X) &= \frac{P_{X^0}(\boldsymbol{X}^0) \delta(||\boldsymbol{X}^0||_0 - NP\rho)}{V_X(\mathcal{Q}_X)} \\ &\times \prod_{a=1}^{n} \delta(||\boldsymbol{X}^a||_0 - NP\theta) \prod_{a\le b} \delta(\mathrm{Tr}(\boldsymbol{X}^a)^{\mathrm{T}} \boldsymbol{X}^b - NPq_X^{ab}), \end{aligned} \qquad (22)$$

where $V_D(\mathcal{Q}_D)$ and $V_X(\mathcal{Q}_X)$ are the normalization constants. The above-mentioned computation provides the following expression:

$$\begin{aligned} &[Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0 \\ &\quad = \int d(NM\mathcal{Q}_D) d(NP\mathcal{Q}_X) V_D(\mathcal{Q}_D) V_X(\mathcal{Q}_X) \\ &\quad \times \left[\prod_{a=1}^{n} \exp\left(-\frac{\beta}{2} \sum_{\mu,l} {t_{\mu l}^a}^2 \right) \right]_{\mathcal{Q}_X, \mathcal{Q}_D}, \end{aligned}$$

where $t_{\mu l}^a = N^{-1/2} \sum_{i=1}^{N} (D_{\mu i}^a X_{il}^a - D_{\mu i}^0 X_{il}^0)$. Notation $[\cdots]_{\mathcal{Q}_D, \mathcal{Q}_X}$ represents the average with respect to $\{\boldsymbol{D}^a\}$ and $\{\boldsymbol{X}^a\}$ within the state space specified by $\mathcal{Q}_D$ and $\mathcal{Q}_X$, whose distributions are given by eqs. (21) and (22). Distributions (21) and (22) are independent of each other, and provide each entry of $\{\boldsymbol{D}^a\}$ and $\{\boldsymbol{X}^a\}$ with zero mean and a finite variance. This allows us to utilize the central limit theorem indicating that we can handle $t_{\mu l}^a$ as multivariate Gaussian random variables that follow

$$P_t(\{t_{\mu l}^a\}) = \prod_{\mu l} \frac{1}{\sqrt{(2\pi)^n \det \mathcal{T}}} \exp\left(-\frac{1}{2} \sum_{a,b} t_{\mu l}^a (\mathcal{T}^{-1})^{ab} t_{\mu l}^b \right), \qquad (23)$$

where $\mathcal{T}$ stands for an $n \times n$ matrix whose entries are given as $\mathcal{T}^{ab} = q_D^{ab} q_X^{ab} - (q_D^{a0} q_X^{a0} + q_D^{b0} q_X^{b0}) + \rho$. Utilizing this and evaluating integrals of $\mathcal{Q}_X$ and $\mathcal{Q}_D$ by means of the saddle point method lead to an expression

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N^2} [Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0 &= \mathrm{extr}\Big[-\frac{\alpha\gamma}{2} \ln\det(\mathcal{I}_n + \beta\mathcal{T}) \\ &+ \gamma \Big\{ \frac{\mathrm{Tr}\hat{\mathcal{Q}}_X \mathcal{Q}_X}{2} + \ln\left(\int \{ \prod_{a=0}^{n} dX^\alpha \} P_{X^0}(X^0) e^{-\Xi} \right) \Big\} \\ &+ \alpha \Big\{ \frac{\mathrm{Tr}\hat{\mathcal{Q}}_D \mathcal{Q}_D}{2} - \frac{1}{2} \ln\det \hat{\mathcal{Q}}_D \Big\} \\ &+ n\lambda\theta + \frac{n\alpha}{2} \ln(2\pi) \Big]. \end{aligned} \qquad (24)$$

Here, $\mathcal{I}_n$ represents the $n \times n$ identity matrix, auxiliary variables $\hat{\mathcal{Q}}_D \equiv (\hat{q}_D^{ab})$ and $\hat{\mathcal{Q}}_X \equiv (\hat{q}_X^{ab})$ are introduced in evaluating $V_D(\mathcal{Q}_\mathcal{D})$ and $V_X(\mathcal{Q}_\mathcal{X})$ with use of the saddle point method, and $\Xi \equiv \frac{1}{2}\sum_{a,b=0}^n \hat{q}_X^{ab} X^a X^b + \lambda \sum_{a=1}^n \lim_{\epsilon\to+0} |X^a|^\epsilon$. Extremization should be taken with respect to λ and four kinds of macroscopic variables $\mathcal{Q}_D$, $\mathcal{Q}_X$, $\hat{\mathcal{Q}}_D$, and $\hat{\mathcal{Q}}_X$.

Exactly evaluating eq. (24) should provide the correct leading order estimate of $N^{-2}\ln[Z_\beta^n(\boldsymbol{D}^0, \boldsymbol{X}^0)]_0$ for each of $n \in \mathbb{N}$. However, we here restrict the candidate of the dominant saddle point to that of the replica symmetric form as

$$q_D^{ab} = \begin{cases} 1, & a = b \\ q_D, & a \neq b,\ (a, b \neq 0) \\ m_D, & a = 0, b \neq 0 \end{cases} \tag{25}$$

$$q_X^{ab} = \begin{cases} Q_X, & a = b \\ q_X, & a \neq b,\ (a, b \neq 0) \\ m_X, & a = 0, b \neq 0 \end{cases} \tag{26}$$

$$\hat{q}_D^{ab} = \begin{cases} \hat{Q}_D, & a = b \\ -\hat{q}_D, & a \neq b,\ (a, b \neq 0) \\ -\hat{m}_D, & a = 0, b \neq 0 \end{cases} \tag{27}$$

$$\hat{q}_X^{ab} = \begin{cases} \hat{Q}_X, & a = b \\ -\hat{q}_X, & a \neq b,\ (a, b \neq 0) \\ -\hat{m}_X, & a = 0, b \neq 0 \end{cases} \tag{28}$$

so as to obtain an analytic expression with respect to n. This yields

$$\begin{aligned} \ln\det(\mathcal{I}_n + \beta\mathcal{T}) &= n\ln(1 + \beta(Q_X - q_D q_X)) \\ &+ \ln\left(1 + n\frac{\beta(q_D q_X - 2m_D m_X + \rho)}{1 + \beta(Q_X - q_D q_X)}\right), \end{aligned} \tag{29}$$

$$\begin{aligned} &\frac{\mathrm{Tr}\hat{\mathcal{Q}}_X\mathcal{Q}_X}{2} + \ln\left(\int\prod_{a=0}^n dX^a P_{X^0}(X^0)e^{-\Xi}\right) \\ &= \frac{n}{2}\hat{Q}_X Q_X - n\hat{m}_X m_X - \frac{n(n-1)}{2}\hat{q}_X q_X \\ &+ \ln\langle\langle\left(\int dX e^{-\xi}\right)^n\rangle\rangle_h, \end{aligned} \tag{30}$$

and

$$\begin{aligned} &\frac{\mathrm{Tr}\hat{\mathcal{Q}}_D\mathcal{Q}_D}{2} - \frac{1}{2}\ln\det\hat{\mathcal{Q}}_D \\ &= \frac{n}{2}\hat{Q}_D - n\hat{m}_D m_D - \frac{n(n-1)}{2}\hat{q}_D q_D \\ &- \frac{n}{2}\ln(\hat{Q}_D + \hat{q}_D) - \frac{1}{2}\left(1 - n\frac{\hat{q}_D + \hat{m}_D^2}{\hat{Q}_D + \hat{q}_D}\right), \end{aligned} \tag{31}$$

where $\xi \equiv (\hat{Q}_X + \hat{q}_X)X^2/2 - hX + \lambda\sum_{\epsilon\to+0}|X|^\epsilon$. Further, the following replacement of variables is convenient in handling our computation in the limit of $\beta \to \infty$: $\hat{Q}_D + \hat{q}_D \to \beta\hat{Q}_D$, $\hat{q}_D \to \beta^2\hat{\chi}_D$, $1 - q_D \to \chi_D/\beta$, $\hat{Q}_X + \hat{q}_X \to \beta\hat{Q}_X$, $\hat{q}_X \to \beta^2\hat{\chi}_X$, $Q_X - q_X \to \chi_X/\beta$, and $\lambda \to \beta\lambda$. In $\beta \to \infty$, integral with respect to X in eq. (30), $\int dXe^{-\xi}$, is replaced to $e^{-\beta\phi(h;\hat{Q}_X,\lambda)}$ by applying the saddle point method. Inserting eqs. (29)–(31) and the rescaled variables into eq. (24) offers the expression of the zero temperature free energy density (3).

REFERENCES

[1] Starck J. -L., Murtagh F. and Fadili J. M., *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity* (Cambridge Univ. Press), 2010.

[2] Nyquist H., *Trans. AIEE*, **47** (1928) 617.

[3] Malviya S., Voepel-Lewis T., Eldevik O. P., Rockwell D. T., Wong J. H., and Tait A. R., *British Journal of Anaesthesia*, **84** (2000) 743.

[4] Lian L. Y., and Robert G. (Eds.), *Protein NMR Spectroscopy: Principal Techniques and Applications* (John Wiley & Sons Ltd.) 2011.

[5] Donoho D., *IEEE Trans. Inform. Theory*, **52** (2006) 1289.

[6] Candes E. J., and Tao T., *IEEE Trans. Inform. Theory*, **51** (2005) 4203.

[7] Kabashima Y,, Wadayama T., and Tanaka T., *J. Stat. Mech.*, (2009) L09003.

[8] Donoho D. L., Maleki A., and Monatanri A., *PNAS*, **106** (2009) 18914.

[9] Ganguli S. and Sompolinsky H., *Phys. Rev. Lett.*, **104** (2010) 188701.

[10] Krzakala F., Mézard M., Sausset F., Sun Y. F., and Zdeborova, L. , *Phys. Rev. X*, **2** (2012) 021005.

[11] Rubinstein R., Bruckstein A. M., and Elad M., *Proc. of IEEE*, **98** (2010) 1045.

[12] Elad M., *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, (Springer-Verlag), 2010.

[13] Gleichman S., and Eldar Y. C., *IEEE Info. Theor.*, **57** (2011) 6958.

[14] Bryt O. and Elad M., *J. Vis. Commu. Image Rep.*, **19** (2008) 270.

[15] Gong T., Xuan J., Chen L., Riggins R. B., Li H., Hoffman E. P., Clarlke R., and Wang Y., *BMC Bioinformatics 2011*, **12** (2011) 82.

[16] Aharon M., Elad M., and Bruckstein A. M., *Linear Algebra and its Applications*, **416** (2006) 48.

[17] Olshausen B. A. and Field D. J., *Vision Res.*, **37** (1997) 3311.

[18] Engan K., Aase S. O., and Hakon Husoy J., *IEEE Acoustic, Speech and Signal Processing*, (1999) 2443.

[19] Aharon M., Elad M., and Bruckstein A. M., *IEEE Trans. Signal Processing*, **2006** (54) 11.

[20] Dotzenko V., *Introduction to the Replica Theory of Statistical Systems* (Cambridge Univ. Press), 2001.

[21] Mézard M., Parisi G., and Virasoro M. A., *Spin Glass Theory and Beyond*, (World Sci. Pub.) 1987.

[22] Guo D. and Verdú S., *IEEE Trans. Inform. Theory*, **51** (2005) 1983.

[23] de Almeida J. R. L. and Thouless D. J., *J. Phys. A: Math. Gen.*, **11** (1978) 983.

[24] Mézard M. and Montanari A., *Information, Physics, and Computation* (Oxford Univ. Press), 2009.

[25] Sakata A., and Kabashima Y., unpublished.